Szövegek értelmezése elnagyolt szintaktikai elemzés alapján

1. B e v e z e t é s . – A szövegek elemzése, automatikus értelmezése nagy kihívást jelent a nyelvészeknek és az informatikusoknak egyaránt. Számos egymással versengő megközelítés jelent meg a gyakorlati megvalósítások elméleti hátterében. A statisztikai és a szabályalapú rendszerek nagyon különböző alapokon próbálják megoldani ugyanazt a feladatot: minél pontosabban behatárolni, hogy miről is szól a számítógépen megjelenő szöveg. A statisztikai módszer előnyben van a szabályalapúval szemben, amikor az utóbbi nem tud pontos elemzést készíteni, akár a szöveg hibái, akár információhiány miatt. Szerencsére a szabályalapú elemzések is módosíthatóak úgy, hogy ezekben a helyzetekben is ki tudjanak nyerni információt a szövegből. Minden bizonnyal számtalan módja van ezeknek a módosításoknak. Ezek közül egyet emelek ki, az elnagyolt szintaktikai elemzést.

Cikkemben[1] egy olyan szemantikai rendszert vázolok fel, amely elnagyolt szintaktikai elemzések eredményét felhasználva törekszik minél pontosabb és megbízhatóbb információkat kinyerni az elemzett szövegből. Mindezt nem a teljes szintaktikai elemzés és egy hozzá tartozó kompozicionális szemantikai szabályrendszer alternatívájának tekintem, hanem ezek melletti kiegészítő módszernek.

2. T e l j e s e l e m z é s a l k a l m a z á s a . – Amikor teljes szintaktikai elemzésre alapozzuk a szöveg értelmezését, a következőkre van szükség:

a) Lexikon, hogy a szöveg alapelemeit fel tudjuk ismerni és a további elemzésekben az itt tárolt információkra (pl. morfológiai vagy szemantikai adatok) tudjunk alapozni.
b) Morfológiai elemző.
c) Szintaktikai elemző.
d) Szemantikai szabályok (valószínűleg a kompozicionalitás elvét követve: minden egyes szintaktikai szabálynak van egy szemantikai megfelelője).
e) Formalizálás minden szinten.
f) Végeredményben egy teljesen formalizált elemzést kapunk a szövegről.

3. A t e l j e s s z i n t a k t i k a i e l e m z é s m e l l e t t f e l l é p ő n e h é z s é g e k . – A szintaktikai elemzés a morfológiai elemzésre épül, ami viszont nem feltétlenül egyértelmű. Hasonlóképpen, több szintaktikai elemzés is lehetséges. A többértelműséget a szemantikai szinten is fel kell oldani (JURAFSKY–MARTIN 2009).

[1] Itt szeretném megköszönni cikkem két lektorának alapos és építő kritikáját. Továbbá köszönöm a hozzászólásokat a Félúton konferencián elhangzott előadásomhoz, valamint a szervezők munkáját, hogy lehetővé tették az előadás megtartását és a cikk megjelenését. Minden igyekezetem és a segítségek ellenére írásomban előfordulhatnak hibák, melyekért a kedves olvasó elnézését kérem.

Mondatszintű elemzésnél pontosabb lehet a szövegszintű. A szöveg szintjén fontos szerepet kap a referenciák kezelése, ami szintén több elemzési lehetőséget eredményezhet, de végeredményben a szövegszintű formális elemző rendelkezésére álló információk lehetőséget adnak az alsóbb elemzési szinteken keletkezett többértelműségek kiszűrésére (SCHULER 2002).

A szabályos elemzések előbb említett nehézségei mellett egyéb problémák is adódnak, ha a teljes szintaktikai elemzésre hagyatkozhatunk csak.

4. A nem ideális szövegek elemzésének problémái. – A valóságban nem biztos, hogy olyan szövegek elemzése lenne a leggyakoribb feladat, melyek ideálisan, hibátlanul vannak megfogalmazva. A teljes szintaktikai elemzés ideális tárgyai a jól szerkesztett mondatok. Ha a mondat szerkezetébe valamilyen hiba csúszik (akár szintaktikai, akár morfológiai vagy akár lexikai szinten), akkor a teljes szintaktikai elemzésre alapuló szemantikai elemzés gyakorlatilag kútba esik. Ugyanis ha nem sikerül teljes szintaktikai elemzést adni a mondathoz, akkor hiába keressük a (nem létező) szintaktikai elemzéshez a szemantikai párját.

Nem áll szándékomban lebecsülni a teljes szintaktikai elemzésre alapuló rendszereket. Az lenne a legjobb, ha mindig ilyeneket használhatnánk. Azonban vannak hibák a nyelvhasználatban. Ezek a hibák pedig nem vezethetnek oda, hogy semmit nem ért meg az elemző az elhangzottakból. Az emberek a zajos környezetben, nehéz körülmények között is megértik, még a kicsit hibás mondatokat is. Kizárásos alapon (sok esetben a világismeretre támaszkodva) fel tudnak oldani referenciákat a szövegben, melyek az elméleti szabályokkal nem magyarázhatók.

Teljes, jól működő szintaktikai elemzőt nehéz építeni. Sok nyelvészeti ismeretet igényel, temérdek adatot fel kell dolgozni, s később ezeket karban is kell tartani. Ennek megkönnyítésén, hatékony rendszerek építésén többen, különböző megközelítéssel is dolgoznak (ALLEN 1995, HÓCZA 2008), de még így is számtalan problémával kell szembenézni. Egy bizonyos nyelvet leírni nem elég. Számtalan nyelvjárást, idiolektust fel kell tudni ismerni, ami az embereknek általában nem okoz problémát.

Ezzel szemben, ha egy idegen nyelvet kezdő szinten beszélünk és a szavaknak, szófordulatoknak csak nagyon kis részét ismerjük, a szintaktikai szabályoknak pedig csak kis hányadát tudjuk, akkor is „elcsípünk” néhány szót ennek a nyelvnek az anyanyelvi beszélőitől, valamit – töredék-információkat – megértünk belőle.

Kis magyar és angol nyelvi ismerettel is megértjük a következőket:

(1) *I not know.*

(2) *Én nem tudni beszélni magyar.*

(3) *Én menni vissza Washington. Mary is.*

5. Töredékes elemzés. – Töredékes elemzés alatt a mondatnál kisebb szintaktikai egységek felismerésén alapuló elemzést értek, melynek eredménye a felismert részek értelmezése, vagyis szemantikai jellegű.

A) E l ő n y ö k . – A töredékes elemzés előnye, hogy az elemzendő szöveg aránylag kis részének sikeres részleges elemzése esetében is kit tudunk nyerni információt a szövegből. Elég, ha egy mondatban felismerjük az alanyi és állítmányi jellegű részeket (vagy azok egy részét), mert ebből már állítás, formula készíthető.

(4) *Mari elment a ~~tegnapi~~ buliba a ~~pasijával~~.*

(5) *Mari elment a buliba.*

Az információk java részének kinyeréséhez elég kisebb részeket felismerni, a hiányok nem feltétlenül végzetesek az elemzés szempontjából. Például a főnévi csoporton belüli jelzős szerkezetekből is ki lehet nyerni információt.

(6) *A szép Mari ~~szombaton elment bulizni~~.*

(7) *Mari szép.*

A töredékes elemzés egyik legnagyobb előnye, hogy a kisebb részek, szerkezetdarabok felismerése technikailag is egyszerűbb. Gondolhatunk itt a véges állapotú automatákkal történő NP-felismerésre is (VÁRADI 2003).

B) N e h é z s é g e k . – Természetesen a részleges elemzést alkalmazó nyelvészeknek, programozóknak is komoly kihívásokkal kell megküzdeniük.

Amíg a teljes szintaktikai elemzés és a neki megfelelő szemantikai elemzés egyértelműen (természetesen csak az egyértelműsítések után, hiszen a többértelműséget addig nem tudjuk kizárni) meghatározza a predikátum–argumentum szerkezetet, egy töredékes elemzés esetében ez külön, komoly feladat. Meg kell határozni, hogy ahhoz a predikátum-jellegű mondatrészhez, amit azonosítottunk, mely azonosított mondatrészek tartozhatnak argumentumként. Elképzelhető, hogy több is van, mint amennyi kellene, de természetesen kevesebb is előfordulhat. Az is lehet, hogy az elemzés töredékes volta miatt túl kevés rendelkezésünkre álló információ következtében egyáltalán nem találunk megfelelő argumentumot hozzá. Mindegyik lehetséges esetet kezelnünk kell ezek közül.

Ha nem találunk megfelelő argumentumjelöltet, akkor nem tudunk állítást megfogalmazni, de legalább biztosak lehetünk benne, hogy nem teszünk téves állítást. Téves alatt nem arra gondolok, hogy hamis az állítás, hanem arra, hogy a predikátum más elemre vonatkozik, mint amire alkalmazzuk. A továbbiakban is fogom használni ezt a kifejezést, ugyanebben az értelemben.

Ha egy argumentumhelyre csak egy azonosított mondatrész a jelölt, és minden argumentumhelyre van jelöltünk, akkor megfogalmazhatunk egy állítást. Természetesen, ha csak részleges elemzést tudtunk végrehajtani, és maradtak elemezetlen részei a mondatnak, akkor az állítás megfogalmazásával kockázatot vállalunk, mert lehet, hogy téves állítást veszünk fel.

A téves állítás megfogalmazása reális veszély, komoly félreértésekhez vezet, ha nem kezeljük ezt a veszélyforrást. Például elképzelhető, hogy a töredékes elem-

zés hiányzó részében maradt egy tagadószó, és így pont ellenkező állítást fogalmazunk meg, mint amit kellene. Hasonlóan komoly félreértést eredményezhet ilyenkor, ha egy összetett kifejezésnek csak egy részét értjük meg.

(8) *Mari elment a buliba. → Mari elment.*

(9) *Mari itta meg a levét. → Mari itta meg. (Mit is?)*

Ne feledkezzünk meg arról az esetről sem, amikor több mondatrész is megfelelőnek tűnhet ugyanarra az argumentumpozícióra, és ilyen módon több állítást is megfogalmazhatunk. Ekkor is fennáll a téves állítás megfogalmazásának veszélye. Ilyenkor segítségünkre lehet a választásban néhány további információ: például a toldalékok, tematikus szerepek (GILDEA– JURAFSKY 2002, PRADHAN–SAMEER–HACIOGLU–WARD–MARTIN–JURAFSKY 2003, SASS 2009, CONNOR–GERTNER–FISHER–ROTH 2010). Ahhoz, hogy ezekre támaszkodhassunk, komoly lexikonra van szükség a háttérben, amely részletes vonzat- és argumentumkeret információkkal rendelkezik. Egy ilyen lexikon felépítése aránylag költséges, mert sok „kézi" munkát kell befektetni.

6. A t é v e d é s e k k e z e l é s e . – Mint láthattuk, számtalan forrásból kerülhet tévedés a rendszerbe. Ha a részleges szintaktikai elemzést használjuk, akkor a tévedések lehető legnagyobb fokú kiküszöbölésére ki kell dolgozni valamilyen módszert.

A tévedések kiküszöbölésének három alappillére van. Egyik a megfelelő struktúra, amely lehetővé teszi két művelet végrehajtását. Másik a tévedés felismerése, a harmadik pedig a téves állítás visszavonása (vagy korábbi állítások felülbírálása). Állítások egy bizonyos elméleten belüli felülbírálásával foglalkozik a belief revision irodalma (ALCHOURRÓN–GÄRDENFORS–MAKINSON 1985). Azzal nem foglalkozik, hogy mit is kellene visszavonni, hanem csak azzal, hogy a visszavonás milyen tulajdonságokkal bír. Mivel kész recept nem áll a rendelkezésünkre, magunknak kell kitalálni az alkalmazható rendszert.

A) Egy b o n y o l u l t , d e a r á n y l a g p o n t o s t é v e d é s - k e z e l ő r e n d s z e r k ö r v o n a l a i . – Röviden összefoglalva így lehetne jellemezni ezt a rendszert: „Mit hogyan mondanánk jól?".

Komoly pragmatikai vizsgálatra lenne hozzá szükség, mert sok apró részletet kellene megvizsgálni. Ha lenne egy alkalmazható pragmatikai elmélet arról, hogy a mondanivalónk megfogalmazásának milyennek kellene lennie, akkor követhetnénk a következő módszert: A feltételezett állításokat, amiket próbálunk megfogalmazni a rendelkezésünkre álló töredékes elemzés alapján, megpróbálnánk belehelyezni a szövegkörnyezetbe, és a pragmatikai elmélet által megjósolt előfordulási valószínűségeket figyelembe véve a legvalószínűbb szerkezetet alkalmaznánk. Esetleg egyéb statisztikai adatokat is figyelembe vehetnénk, mint például a szintaktikai szerkezetek gyakorisága: milyen szintaktikai szerkezeteket szoktunk alkalmazni; amit most alkalmaznánk, az milyen valószínűséggel fordulhat elő ebben a környezetben.

Ehhez természetesen végig figyelemmel kellene kísérni a szövegben (többek között) a régi és új információk megjelenését is.

Ha az állítások sorrendjének a valószínűségei rendelkezésünkre állnak, remélhetőleg kellően nagy biztonsággal meg tudjuk mondani, hogy mely állítás téves és melyik nem, és az utóbbiak közül melyik a legelfogadhatóbb.[2]

B) Egy szerényebb eszközökkel is megvalósítható módszer a tévedések kezelésére: ellentmondások vizsgálata a szövegben. – Ehhez a módszerhez néhány alapvető műveletre és tevékenységre van szükség. A szöveg vizsgálata során folyamatosan figyelni kell, hogy az új információ ellentmond-e az eddigieknek. Ha ellentmond, vissza kell vonni az új állítást vagy a tagadását. Élő helyzetben megvan az a lehetőségünk, hogy visszakérdezzünk, melyiket kell visszavonnunk a kettő közül. Szöveg olvasásakor vagy más, korlátozottabb környezetben a kevésbé megbízható állítást kell visszavonnunk.

Amit most felsoroltam, valóban csak néhány egyszerű művelet, és néhány nagyon egyszerű tevékenység, melyek automatizálhatóak. Össze sem hasonlítható ennek a rendszernek a bonyolultsága egy részletes pragmatikai elmélet összetettségével.

Ha ezt a módszert követjük, akkor rövid szövegekben kisebb, hosszú szövegekben nagyobb biztonsággal ki tudjuk küszöbölni a tévedéseket utólag. (Valószínűleg egy hosszabb szövegben az egyes szövegrészek vagy megismétlődnek, vagy előkerülnek korábbi formulák következményei, ezért ellentmondás merül fel. Igaz, hogy az ellentmondás előfordulása már nagyon erős követelmény, de megléte biztosítja, hogy felismerjünk egy téves formulát. Szerencsére – igaz, hogy kisebb biztonsággal –, de más adatokra is támaszkodhatunk.)

A továbbiakban ennek a viszonylag egyszerű módszernek a pontosabb ismertetésével foglalkozom.

7. Megbízhatóság. – Ahhoz, hogy pontosan értelmezni tudjuk a fent leírt műveleteket, szükségünk van a megbízhatóság fogalmára, mert össze kell tudnunk hasonlítani a formulák megbízhatóságát. Az alábbiakban több fajta megbízhatóságot is megkülönböztetek.

A) Formulák egyéni megbízhatósága. – A formulák megbízhatóságának pontos definiálásakor a következő tulajdonságokat kell figyelembe vennünk:

A teljesen elemzett mondathoz tartozó formula megbízhatóbb, mint a részlegesen elemzetthez tartozó.

A teljesen elemzett mondatrészből származó töredék-információ ugyanolyan megbízható, mint a teljesen elemzett mondatból származó. (Itt például a főnévi cso-

[2] Ezekben a bekezdésekben leginkább szabályalapúként jellemezhető rendszert vázoltam, statisztikai támogatással. Cikkem nagyra becsült lektorának igaza van abban, hogy ez teljesen statisztikai alapokon, tanító adatokkal, gépi tanulási módszerekkel megvalósítható lenne. Ez teljesen eltérő megközelítés, mint amit ismertetek.

porton belül talált jelzős szerkezetekre gondolok: ha a főnévi csoportot sikerült teljesen elemezni, akkor a belőle kinyert információ megbízhatósága független attól, hogy a teljesen elemzett főnévi csoport részben, vagy teljesen elemzett mondatban szerepel-e. *A szép Mari* mindegy, hogy táncol, vagy játszik, ezektől függetlenül szép.)

Ahhoz, hogy számolni is lehessen a megbízhatósággal és be lehessen építeni egy algoritmusba, érdemes definiálni egy mérőszámot hozzá.

Az előző néhány bekezdésben leírt megbízhatósági értéket a formula elemzési megbízhatóságának nevezem, mert értéke kiszámítható pusztán annak a mondatnak az elemzési módjából, amelynek elemei alapján megfogalmaztuk.

A következőt javaslom a kiszámításra (jelöljük M e l -el az elemzési megbízhatóságot, A -val pedig a formulát, ahogy a későbbiekben is):

$$Mel(A) = Sn/N$$

S n jelöli a szerkezetileg azonosított szavak számát: azoknak a szavaknak a száma, amelyek a legkisebb olyan szerkezetben szerepelnek, amely önálló egységet alkot a mondaton belül. N az összes szó száma, amely ebben a szerkezetben szerepel. Vagyis ha egy főnévi csoporton belüli jelzős szerkezetet teljes egészében megértettünk és minden elemét felhasználva alkottunk egy állítást, akkor ennek az elemzési megbízhatósága ugyanúgy 1, mintha teljes elemzést adtunk volna a mondathoz és abból állítottuk volna össze a formulát.

Az így definiált mérőszám 0 és 1 közötti értéket ad eredményül, és minél nagyobb a szám, annál megbízhatóbbnak tekintjük a formulát.

Nézzünk egy példát a könnyebb érthetőség kedvéért!

(10) *A gyönyörűen éneklő szép Mari felvételizett a főiskolára.*

Ebben a mondatban egyetlen egységnek tekinthetjük azt, hogy *A gyönyörűen éneklő szép Mari*. Ha ez alapján azt az állítást alkotjuk, hogy *Mari szép*, melyet formalizálhatunk például így: $S(m)$, akkor a fentebb említett képlet elemei így alakulnak: $A = S(m); Sn = 5; N = 5$. De ha nem értjük meg azt a szót, hogy *éneklő*, akkor már $Sn = 4$.

Természetesen vitatható, hogy mit tartunk a legkisebb, önálló egységet alkotó szerkezetnek. Minél nagyobb egységekkel dolgozunk, annál megbízhatóbb lesz az elméletünk, de annál sérülékenyebb a szintaktikai elemző. A döntést gyakorlati szempontok is befolyásolhatják: milyen elemekre milyen minőségű elemző áll rendelkezésünkre. A legfontosabb két szempont a nagy megbízhatóság és a kis sérülékenység. (Egy elemző annál sérülékenyebb, minél több hibás vagy zajos bemenő adathoz nem ad eredményt.)

B) F o r m u l á k f o r r á s a i . – Ahhoz, hogy még pontosabban megbecsülhessük a formulák megbízhatóságát, nyilván kell tartani a formulák forrását is. A forrásoknak nemcsak a listáját kell karbantartani, hanem a forrásokhoz is kell rendelni megbízhatóságot. Ennek pragmatikai okai vannak, de nem kell komoly pragmatikai

elméletet keríteni a számítás köré ahhoz, hogy jól és hatékonyan működjön az algoritmusunk.

Mindannyian tapasztaltuk már, hogy vannak nagyon megbízható és nagyon megbízhatatlan információforrások. Például egy sokat tréfálkozó, vagy gyakran téves információkat nyújtó ember megnyilvánulásait nem feltétlenül vesszük komolyan; ugyanígy lehet, hogy komolyabban vesszük a Wikipédia oldalakon leírtakat, mint ugyanabban a témakörben egy kétes vitafórumon olvasható sorokat. A „zajos" vagy problémás forrásokra nem adunk annyira.

A könnyebb számíthatóság érdekében érdemes a források megbízhatósági értékét is egy 0 és 1 közötti számként definiálni.

C) A megbízhatóság számítása. – Valószínűleg lehetne találni még további elemeket is, amelyek befolyásolják egy formula megbízhatóságát, de már a most felsoroltak is komoly előrelépést jelentenek egy formula eredő megbízhatóságának kiszámításában.

Ahhoz, hogy az eredő megbízhatóságot pontosan definiáljuk, előbb érdemes összefoglalni, hogy milyen tulajdonságokkal kell bírnia az eredményt kiszámító képletnek.

Tudjuk azt, hogy a megbízhatóbb forrás(ok)ból származó formulák megbízhatóbbak. Ha a formula több forrásból származik, akkor megfontolandó, hogy melyik forrás megbízhatósági értékét hogyan vegyük figyelembe. Véleményem szerint itt a leggyengébb láncszem elvét kell követni, vagyis a legkevésbé megbízható forrás megbízhatósági együtthatójával kell számolni, hiszen ezen a bizonytalanságon nem javít az, hogy a többi forrás megbízhatóbb.

Ilyen módon a formula eredő megbízhatósági értékének [Mer(A)] kiszámítására a javaslatom:

$$Mer(A) = min(Me(F(A))) * Mel(A)$$

Ahol min(Me(F(A))) a formula forrásai (F(A)) közül a minimális megbízhatósági együttható (Me) és Mel(A) a formula egyéni (elemzési) megbízhatósági együtthatója.

D) A források megbízhatósági értékének karbantartása. – Számtalan esetben kiderülhet, hogy az a forrás, amit eddig megbízhatónak tartottunk, valójában rendszeresen félrevezet minket. Ezért szükséges, hogy ne csak nyilvántartsuk a források megbízhatóságát, hanem tartsuk karban is őket, frissítsük az értéküket, ha kell.

A módosításnál figyelembe kell vennünk a tapasztalatainkat a szóban forgó forrásról. Vagyis, ha kiderül egy formuláról, hogy téves vagy hamis, és vissza kell vonnunk, akkor a formula forrásának vagy forrásainak megbízhatóságát csökkentenünk kell. Természetesen, ha az elemzésünk annyira töredékes volt, hogy gyakorlatilag használhatatlan formulát kaptunk, majd emiatt kellett visszavonni, akkor különös lenne a formulának a forrásaitól független megbízhatatlanságát a forrásain

számon kérni. Ezért minél kisebb a visszavont formula elemzési megbízhatósága, annál kevésbé kell csökkenteni a visszavonás miatt a források megbízhatóságát.

Képletben megfogalmazva:

$$Me(F(A)) = Me(F(A)) * k * (1 - Mel(A))$$

Itt k egy 0 és 1 közötti értéket képviselő, tapasztalati úton beállítandó konstans.

E) A megbízhatóság növelése. – Bizonyos esetekben, amikor eleve bizonytalanok vagyunk abban, hogy jól hallottuk-e, amit hallottunk, vissza szoktunk kérdezni, hogy jól értettük-e, amit mondtak nekünk. Főleg, ha nagyon hihetetlennek tűnik, amit megértetünk belőle, vagy nem fér össze korábbi információinkkal. Esetleg egyenesen ellentmond nekik. Ha ezt a megbízhatósági értékek számértékeire fordítjuk, akkor azt is mondhatjuk, hogy ha egy formula elemzési vagy eredő megbízhatósági értéke egy bizonyos határ alá esik, vagy ellentmondást tapasztalunk, akkor visszakérdezünk. A visszakérdezés annyit jelent, hogy megkérdezzük a beszélőt: igaz-e az állítás, amit leszűrtünk?

A visszakérdezés eredménye több fajta is lehet. Egyrészt megerősítheti a formula megbízhatóságát a beszélő (ha igennel felel, akkor az elemzési megbízhatóságát 1-re állíthatjuk), másrészt gyengítheti is a megbízhatóságát (ha a beszélő nemmel felel, akkor 0-ra állítjuk az elemzési megbízhatóságát, vagyis vissza kell vonnunk, ha már beemeltük az információhalmazunkba, ha pedig nem, akkor be sem szabad emelnünk), továbbá egyéb információkat is kaphatunk (a beszélő válasza eredményezhet ellentmondást és visszakérdezést is).

Elképzelhető, hogy akár a szövegben, akár a visszakérdezés során újra előkerül egy állítás, amely már szerepelt korábban. Ez megerősíti az állítás megbízhatóságát, de korlátozottan. Egyszerűen annyit kell tennünk, hogy a formula elemzési megbízhatóságának értékét az előfordulások közül legmagasabb elemzési megbízhatósági értékkel rendelkező előfordulás értékére állítjuk.

8. Töredékes elemzésre alapuló rendszer moduljai

A) Információábrázolási modul. – Ezt dinamikus szemantikai elméletek (KAMP 1981, KÁLMÁN–RÁDAI 2001) nyomán érdemes definiálni, mert azok képesek a mondathatárokon túlnyúló elemzésre, a teljes szöveg összefüggéseit vizsgálni, kezelni az anaforikus viszonyokat, diskurzus-referenseket. Bele kell vonni a formulák megbízhatósági értékeit és a források nyilvántartását is, ezek nem okoznak komoly gondot.

B) Ellentmondásokat felismerő modul. – Ennek a modulnak folyamatosan ellenőriznie kell, nem került-e ellentmondás a rendszerbe. Ha igen, akkor azonosítania kell az érintett formulákat.

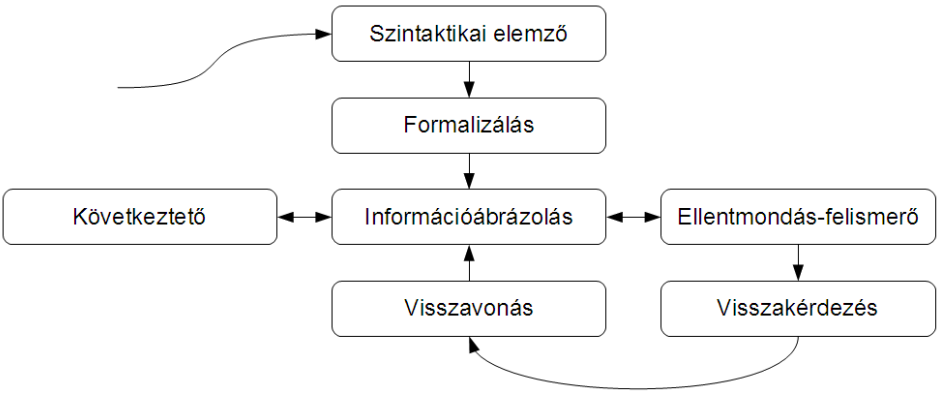
C) Következtető modul. – Érdemes egy ilyen modult is beépíteni egyszerű következtetések levonására. Ennek a modulnak a célja, hogy a hétköznapi

helyzetekben használt, gyakorlatilag bárki által könnyen átlátható és gyorsan elemezhető, deduktív következtetéseket az elemző rendszer is felismerje és használja. Ilyen jellegű például az általános állításokból egyediek levonása. Természetesen, mivel új formulákat vezet le, a források kezelésével is foglalkoznia kell.

D) Visszavonási algoritmus. – Ezen algoritmust megvalósító modulra is szükség van, mert nélküle nem tudjuk kiküszöbölni az ellentmondásokat.

E) Visszakérdezés kezelése. – Erre a modulra akkor van szükség, amikor lehetőségünk van a beszélővel való interakcióra, ellenőrző kérdéseket tudunk feltenni neki, hogy a visszavonás minél hatékonyabban működjön. Szüksége van bemenetként az ellentmondás-ellenőrző adataira. Megvalósítását a kíváncsi szemantika (inquisitive semantics) alapján javaslom (GROENENDIJK–ROELOFSEN 2009).

1. ábra
A modulok kapcsolódása



9. Az ismertetett rendszer gyakorlati (számítógépes) előnyei. – Ha nem áll rendelkezésünkre elég jó szintaktikai elemző, de elérhető laza elemző (shallow parser) összetevőkre, jobb eredményeket kaphatunk. Szintaktikailag helytelen mondatok esetén töredék információkat tudunk kinyerni ahelyett, hogy semmit sem tudnánk meg. Pontatlan beszédfelismerő rendszerek esetén lehetőségünk van a téves felismerések lehetséges kiküszöbölésére szemantikai szinten (ezek a rendszerek még csak szűk szakterületeken használhatók jól).

10. Összefoglalás. – Cikkemben egy olyan keretrendszer vázlatát ismertettem, amellyel kezelni lehet a töredékes elemzésből fakadó problémákat. Igaz ugyan, hogy a töredékes elemzés szintaktikailag könnyebbséget jelent, de a szemantika oldalán bonyolítja a helyzetet, ugyanis a predikátum–argumentum viszonyok

felismerése nem adott, és az ebből, valamint a töredékes elemzés más hiányosságaiból fakadó problémák miatt kiemelt jelentőségű a rendszer hibakezelési módszere. (Hibakezelési módszer egyébként is jó, ha van a teljes szintaktikai elemzés mellett is, nem ennek a módszernek a kifejlesztése és használata okozza a nehézséget.) Az ismertetett rendszernek nagy előnye a teljes szintaktikai elemzéshez képest, hogy a töredékes elemzésből is képes információkat kinyerni, vagyis töredékinformációkat is felismer. Cikkem a szükséges alapmódszereket és elemeket tartalmazza, továbbfejlesztése lehetne a konkrét, formális definíció.

A hivatkozott irodalom

ALCHOURRÓN, CARLOS EDUARDO – GÄRDENFORS, PETER – MAKINSON, DAVID 1985. On the logic of theory change: Partial meet contraction and revision functions. Journal of Symbolic Logic 50: 510–30.

ALLEN, JAMES F. 1995. Natural Language Understanding, The Benjamin/Cummings Publishing Company, Menlo Park, California.

CONNOR, MICHAEL – GERTNER, YAEL – FISHER, CYNTHIA – ROTH, DAN 2010. Starting From Scratch in Semantic Role Labeling. In: ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics Stroudsburg, PA, USA: 989–98.

GILDEA, DANIEL – JURAFSKY, DANIEL 2002. Automatic Labeling of Semantic Roles. Computational Linguistics Volume 28, Issue 3: 245–88.

GROENENDIJK, J. – JANSSEN, T. M. V. – STOKHOF, M. eds. 1981. Formal Methods in the Study of Language. Mathematisch Centrum, Amsterdam.

GROENENDIJK, JEROEN – ROELOFSEN, FLORIS 2009. Inquisitive Semantics and Pragmatics. In: Proceedings of the International Workshop on Semantics, Pragmatics and Rhetorics. Donostia, Spain.

HÓCZA ANDRÁS 2008. A magyar nyelv automatikus szintaktikai elemzése szabályalapú gépi tanulási technikák alkalmazásával. Kéziratos PhD-disszertáció.

JURAFSKY, DANIEL – MARTIN, JAMES H. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition and Computational Linguistics. 2nd edition. Prentice-Hall, USA, New Jersey.

KÁLMÁN LÁSZLÓ – RÁDAI GÁBOR 2001. Dinamikus szemantika. Osiris Kiadó, Budapest.

KAMP, HANS 1981. A Theory of Truth and Semantic Representation. In: GROENENDIJK–JANSSEN–STOKHOF eds. 1981: 277–322.

PRADHAN, SAMEER – HACIOGLU, KADRI – WARD, WAYNE – MARTIN, JAMES H. – JURAFSKY, DANIEL 2003. Semantic Role Parsing: Adding Semantic Structure to Unstructured Text. In: Proceedings of the Third IEEE International Conference on Data Mining (ICDM-2003), Melbourne, FL, Nov.: 19–22.

SASS BÁLINT 2009. Korpusznyelvészeti eszköz a magyar igék bővítményszerkezetének vizsgálatára. In: SINKOVICS szerk. 2009: 143–55

SCHULER, WILLIAM 2002. Interleaved semantic interpretation in environment-based parsing. Proceeding COLING '02 Proceedings of the 19th International Conference on Computational Linguistics. Volume 1. Association for Computational Linguistics Stroudsburg, PA, USA: 1–7.

SINKOVICS BALÁZS szerk. 2009. LingDok 8. Nyelvész-doktoranduszok dolgozatai. JATEPress, Szeged.

VÁRADI TAMÁS 2003. Shallow Parsing of Hungarian Business News. In: Proceedings of the Corpus Linguistics 2003 Conference, Lancaster: 845–51.

DYEKISS EMIL GERGELY (emil.dyekiss@gmail.com)

Shallow Parsing Based Semantic Analysis of Texts

In this article I suggest shallow parsing as a tool for automatic understanding of texts, additionally to full syntactical parsing. Full syntactic parsing of a text helps the exact semantic parsing. If we cannot do full syntactical parsing, but we have shallow parsing and partial results, we can also make formulae for semantic parsing, but with a risk of misparsing. The risk needs to be handled. This can be done by calculating and maintaining the reliability of the formulas and their sources. Catching misformulation can be done by recognizing contradictions caused by adding mistaken formulae. After contraction of one of the formulas of the causes of this contradiction, the less reliable formula is dropped from the formulation of the text. The sketch of this theory can help automatic understanding and information extraction from texts by using shallow parsing.

DYEKISS, EMIL GERGELY